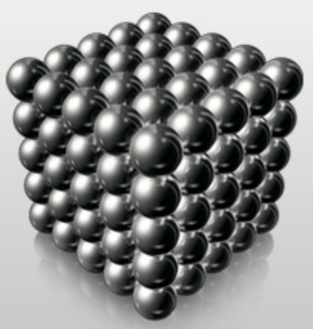
Large Installation and HPC Systems Administration

Session 618

These are confidential sessions—please refrain from filming or taking pictures.



Enterprise IT

Large Installation and HPC Systems Administration

Joshua Durham and Dave Lopata

http://developer.apple.com/wwdc2005/

Agenda

- Cluster computing review
- Architecting an appropriate cluster
- Physical deployment considerations
- Building a basic cluster framework—getting to LINPACK
- Testing the cluster components
- Tweaking OS X for performance
- Job scheduling and resource management

Cluster Computing Review

http://developer.apple.com/wwdc2005/

Available or Scalable Clusters

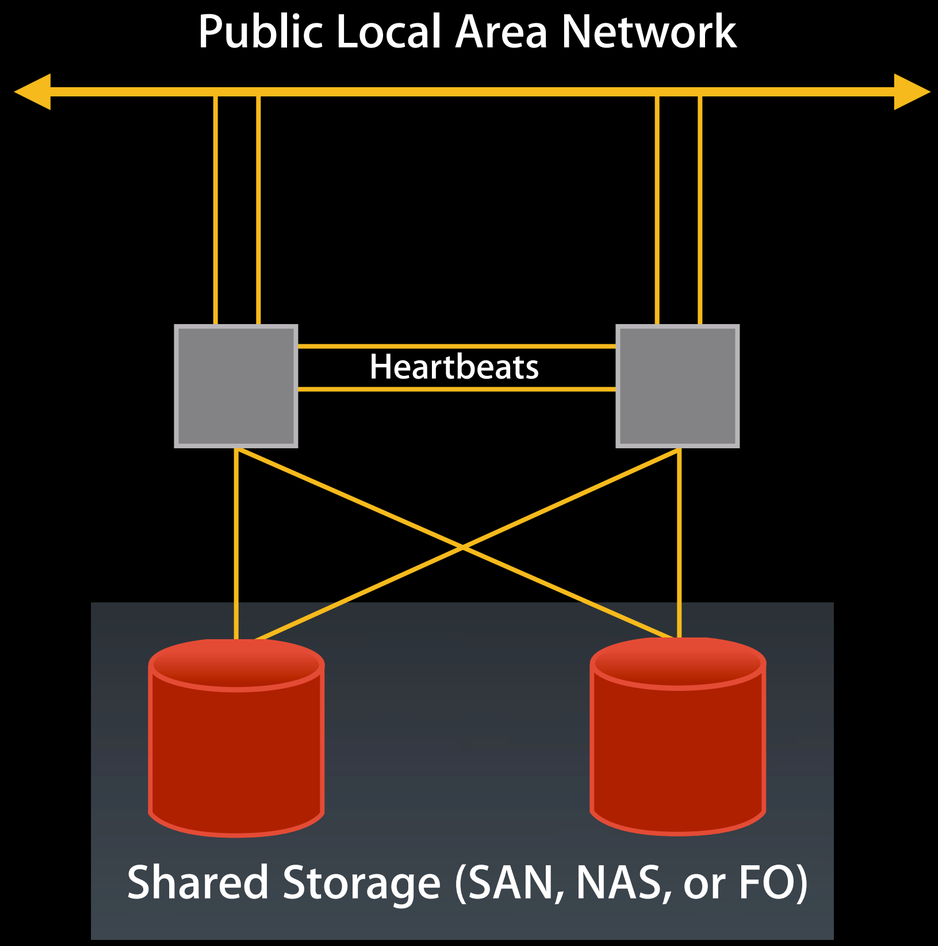
Will Not Be Covered in this Session

http://developer.apple.com/wwdc2005/

High Availability

Application and box-level fail over

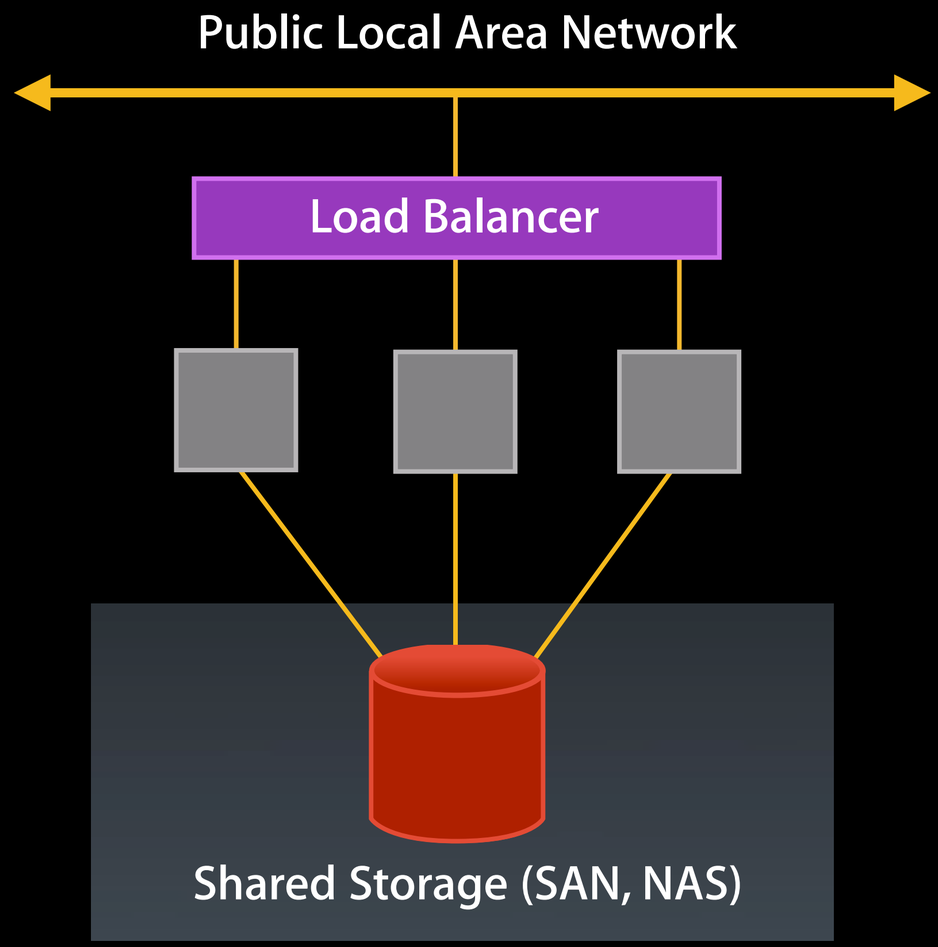
- Typically requires two private networks to act as “heartbeats”
- Typically requires two public network connections per host
- Boot disk is typically mirrored internally
- Data disks are external or remote
- Applications must be cluster aware (i.e. Oracle) or static (i.e. basic web)



Scalable Clusters

Application scalability across multiple systems

- Clustering occurs at the balancer or application level
- Individual hosts are shared state or stateless systems
- Boot disk is typically mirrored internally
- Data disks are external or remote
- Applications must be cluster aware (i.e. Oracle) or static (i.e. basic web)



Computational Clusters

http://developer.apple.com/wwdc2005/

Types of Computational Clusters

High-performance computing

- Research areas like physics, fluid, and molecular dynamics
- Often uses internally-developed software applications (Fortran)
- Typically intensive floating-point calculations
- The computer is the solution

Types of Computational Clusters

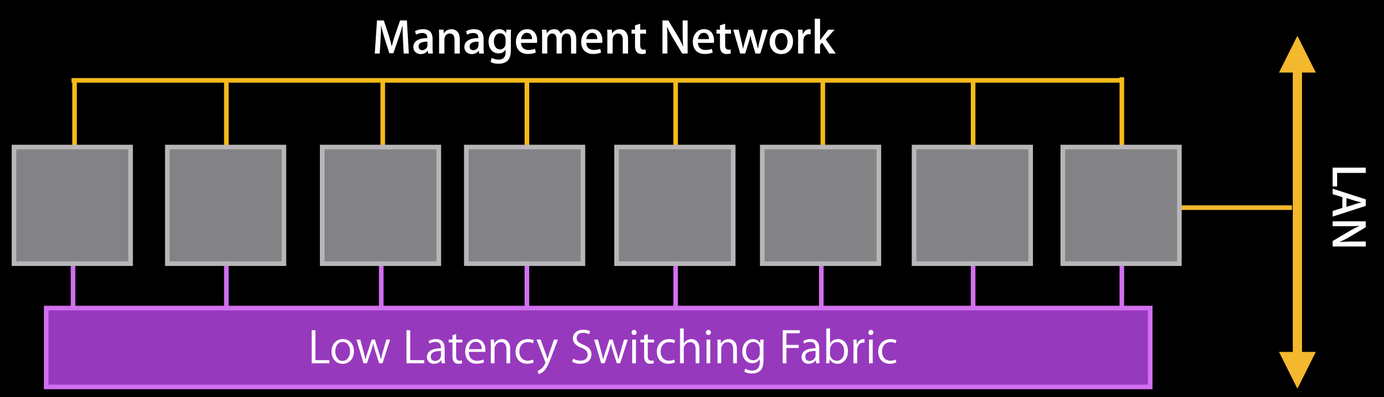
Solutions-based cluster computing

- Biology, agriculture, mathematics, statistics, film schools
- A cluster designed for specific set of known problems
- Solution examples: InQuiry, gridMathematica, Renderman, Xcode
- The software drives the solution

High-Performance Cluster Computing

The computer is the solution

- Generalized needs
  - Good double-precision floating-point performance
    - High-performance LINPACK benchmarks
  - High-speed interconnect
    - Myrinet, Infiniband
- High-performance compilers
  - IBM XL, Absoft



Solutions-Based Cluster Computing

The software is the solution

- Generalized needs
  - Ease of use
    - Quick setup
  - A “product” not a “project”
    - Good documentation and support
- Strong performance
  - Focuses on time to results

INQUIRY

PIXAR's RenderMan®

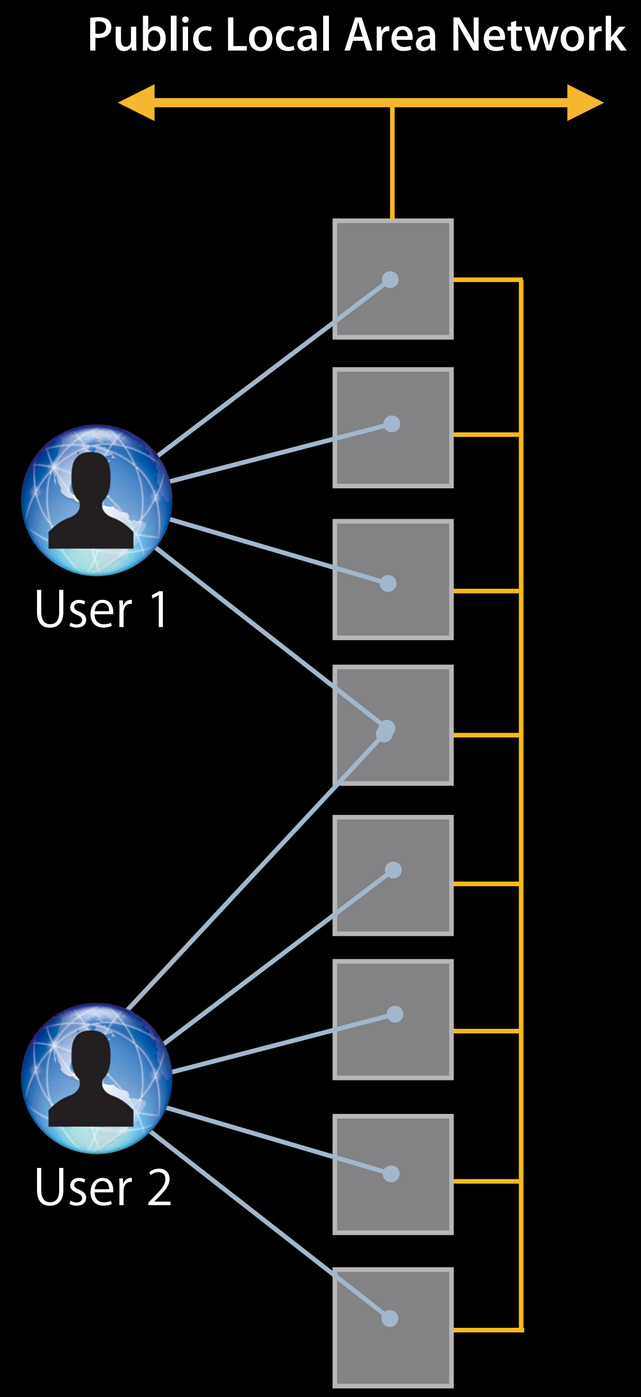
Architectures

http://developer.apple.com/wwdc2005/

Cluster Computing

Ad-hoc or LAN-based computational clusters

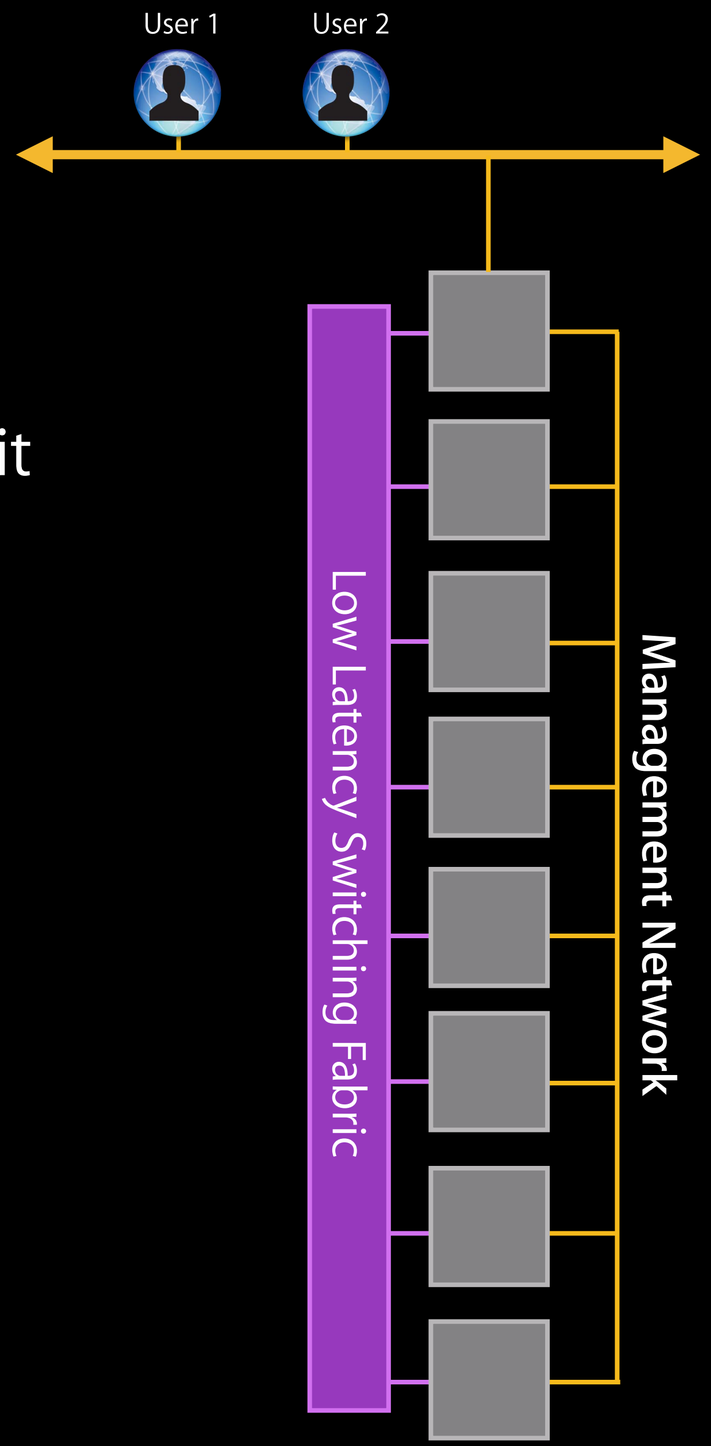
- All nodes sit on the LAN
  - Users login where they see fit
  - Often nodes are dual purpose (workstations)
- Minimal resource management
  - Cluster utilization is often suboptimal
- Single network topology
  - All communications generally over Ethernet



Cluster Computing

Dedicated HPC or research cluster

- Single node (head node) has access to the LAN
  - Users login to head node, or submit jobs through it
- Private service network (Ethernet)
  - Software updates
  - Individual node access
  - Job scheduler interactions
- Dedicated communications network
  - Myrinet, Infiniband
    - Low latency, high speed
  - Ethernet
    - Cost-effective if latency and communications speed is not a factor



Architecting an Appropriate Cluster

http://developer.apple.com/wwdc2005/

Building to an Identified Goal

Future Plans	Fixed purpose, general use, etc
Performance	2.0 Teraflops on LINPACK
Application	HP LINPACK, BLAST, etc
Middleware	MPI, PVM, BEEP
Interconnect	GigE, Myrinet, Infiniband, Firewire
Storage Needs	Xserve, Xserve RAID
Platform	Xserve G5 and OS X

Interconnects
Deciding factors

- Price
  - Gigabit Ethernet is by far the cheapest solution
    - No additional HBAs required
    - Gigabit Ethernet switches are more flexible for later redeployment
- Performance
  - Latency
    - Both Myrinet and Infiniband provide very low latency
  - Throughput
    - Myrinet advertises up to 900 MB/s with dual ported HBAs
    - Infiniband advertises up to ten Gb/s per port on a dual-ported HBAs
      - Note that this greatly exceeds our PCI-X throughput

Physical Deployment Considerations

http://developer.apple.com/wwdc2005/

Deployment

Racks

- Xserve (two- and four-post 19” racks and cabinets)
  - Two-post Telco racks
    - Center mounted
    - Mounting brackets should be installed “flap back”
  - Four-post racks and cabinets
    - Open racks 24- or 26-inches deep (Xserve will extend out the back)
    - Racks or cabinets 29–36-inches deep
- Xserve RAID (four-post 19” racks and cabinets only)
  - Racks or cabinets 24–36-inches deep

Deployment

Racks—suggestions

- Mount order
  - Mount any UPSes at the bottom of the rack
  - Mount Xserve RAID above UPSes, but below servers (fill bottom up)
  - Mount Xserves above RAID and UPSes
- Tools and parts
  - Size zero and one (i.e., small and medium) phillips screw driver needed
  - Xserve comes with all brackets and screws needed
  - Use the screws provided with the Xserve products for attaching all brackets to the product
  - Use the screws provided with the rack for attaching all brackets to the rack

Deployment

Power and environmental—see Kbases 86694 and 86250

Xserve Single CPU	Xserve Dual CPU	Xserve Cluster Node	Xserve RAID
Rated 5.5 A @ 400 W 155–210 W 529–717 BTU/h	Rated 5.5 A @ 400 W 190–290 W 649–990 BTU/h	Rated 5.5 A @ 400 W 161–252 W 550–860 BTU/h	Rated 7.6 A @ 450 W 300–400 W ~1365 BTU/h

Deployment

Power and environmental—see Kbases 86694 and 86250

32 Node HPC Cluster	64 Node Cluster	256 Node Cluster	1110 Node Cluster
One Head Node	One Head Node	One Head Node	Four Head Nodes
One Storage Node	One Storage Node	One Storage Node	Six Storage Nodes
One Xserve RAID	One Xserve RAID	Two Xserve RAIDs	Six Xserve RAIDs
30 Compute Nodes	62 Compute Nodes	254 Compute Nodes	1100 Compute Nodes
~ 8500 Watts	~ 16600 Watts	~65500 Watts	~ 310 KiloWatts
~ 29000 BTU/h	~ 56400 BTU/h	~ 224000 BTU/h	< two million BTU/h

Building a Basic Cluster Framework

Getting to LINPACK

http://developer.apple.com/wwdc2005/

Service and Head Node Setup

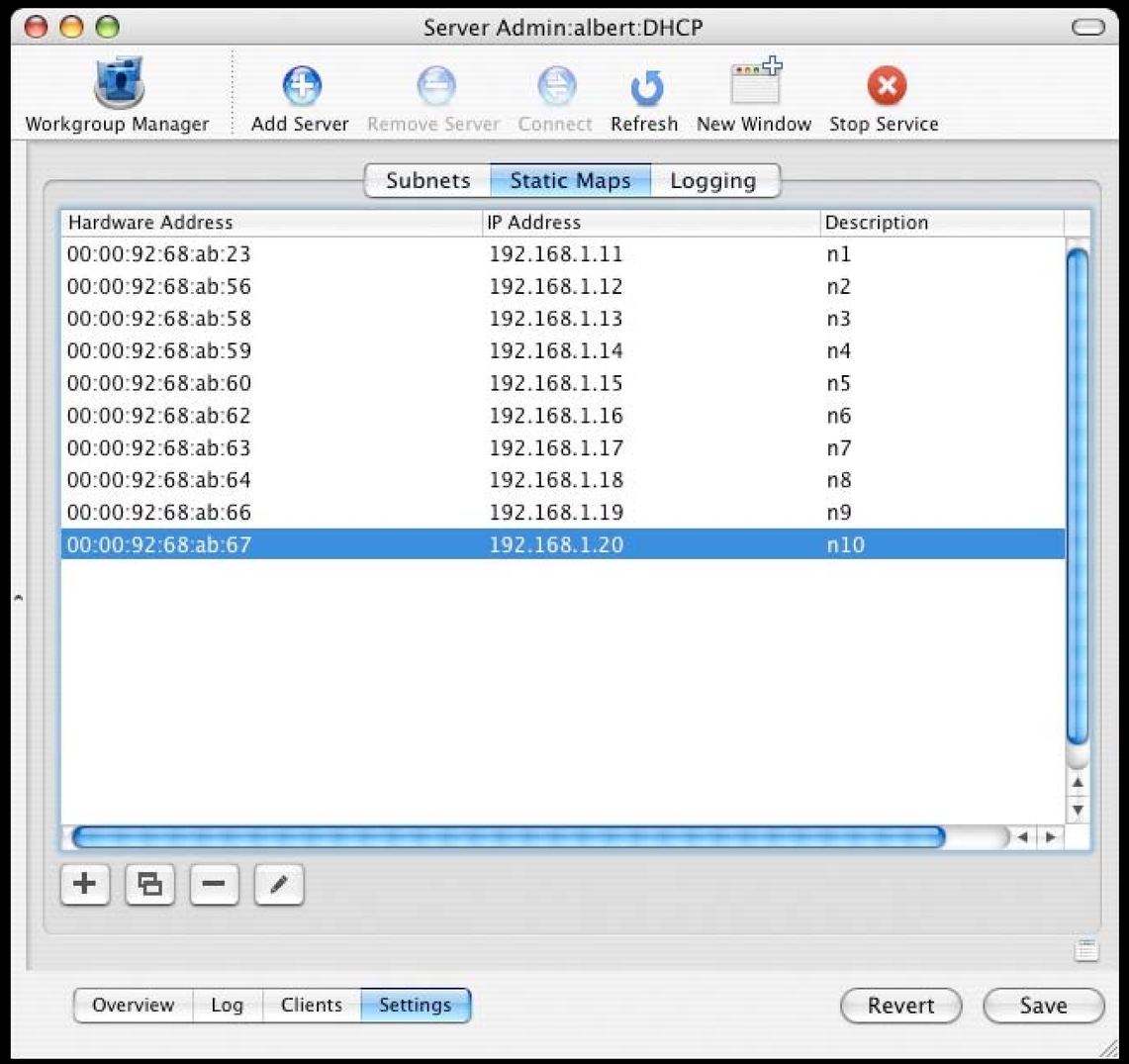
Overview

- A service node runs ‘infrastructure services’ for the cluster
  - DHCP and DNS
  - Common user directory (open directory)
  - Shared file storage (NFS)
  - System hardware monitoring
  - Job queueing and scheduling (if desired)
- Can be one system on smaller clusters
- On larger clusters, consider distributing the services across systems

Service Node Setup

Initial operating system install

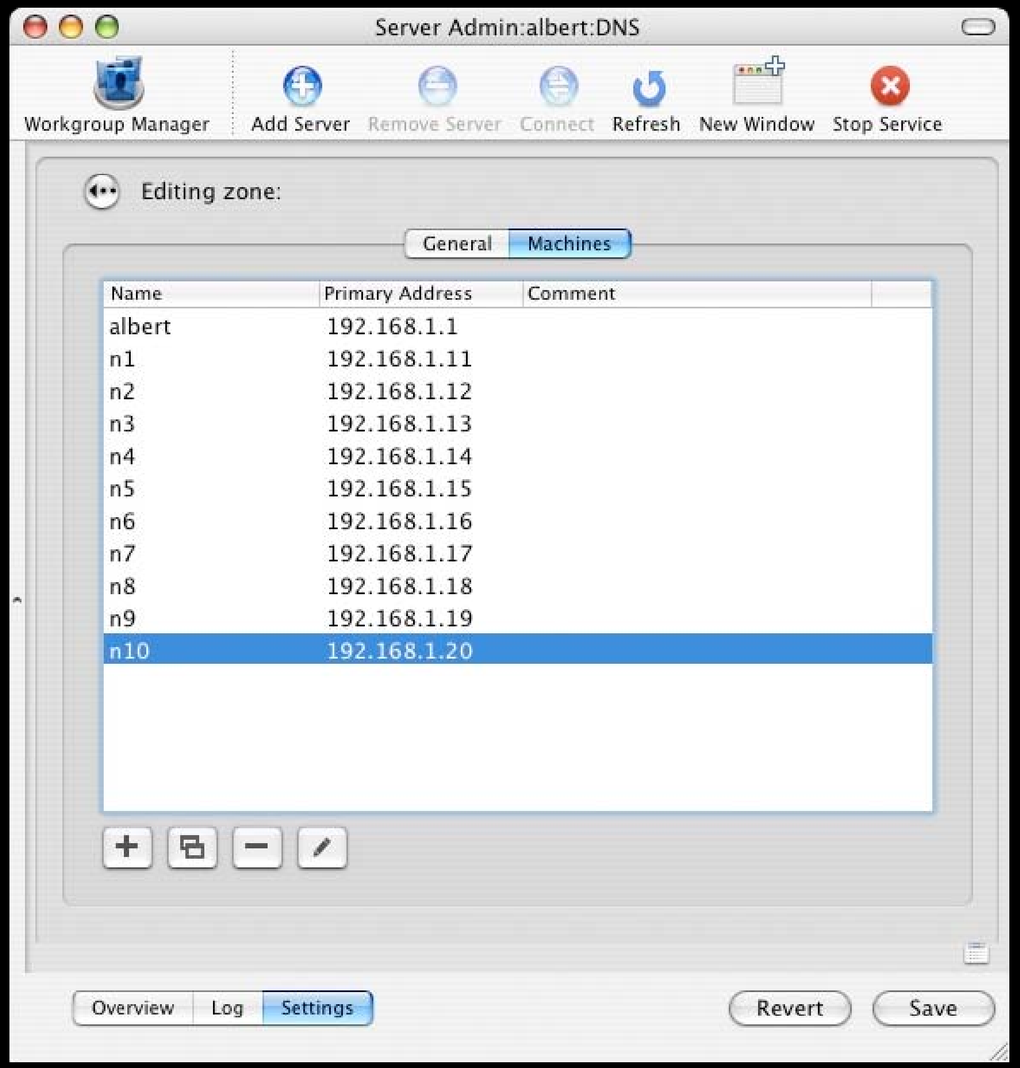
- Installation methods
  - Locally with CD using a keyboard, mouse, and monitor
    - Best for first time (first head node) setups
  - On a laptop (boot Xserve into target disk mode and attach to laptop)
  - Remotely with CDs and server assistant
  - Network install
- Full install of OS X server software



Service Node Setup

DHCP

- DHCP is used to give each node their address
- Psuedo-dynamic IP address allocation
  - We want each compute node to keep their IP address for diagnostic reasons
- Set up using server admin



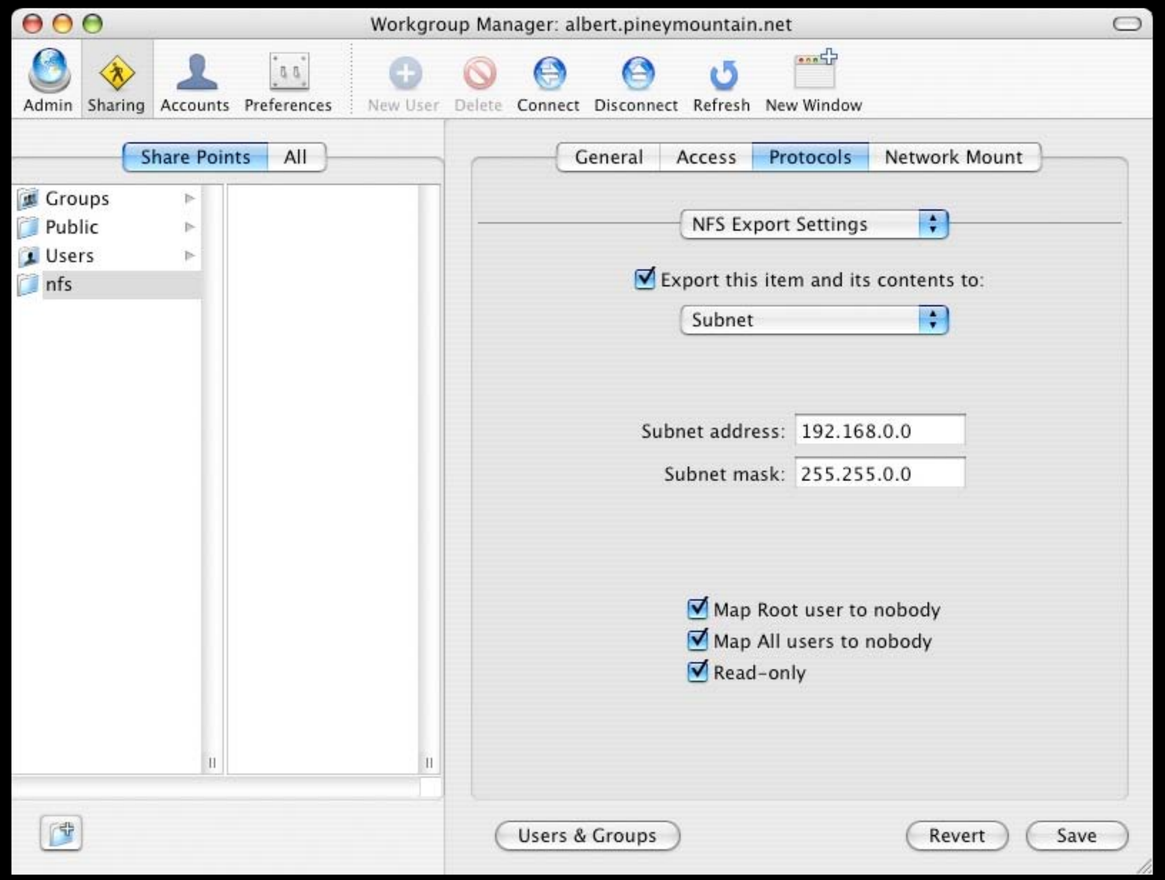
Service Node Setup

DNS

- DNS is used as a central store for hostnames
- Compute nodes should have nicely parsable, reg-ex friendly names
  - Good: node 1–999
  - Bad: node001–node999
- Set up using server admin

Service Node Setup

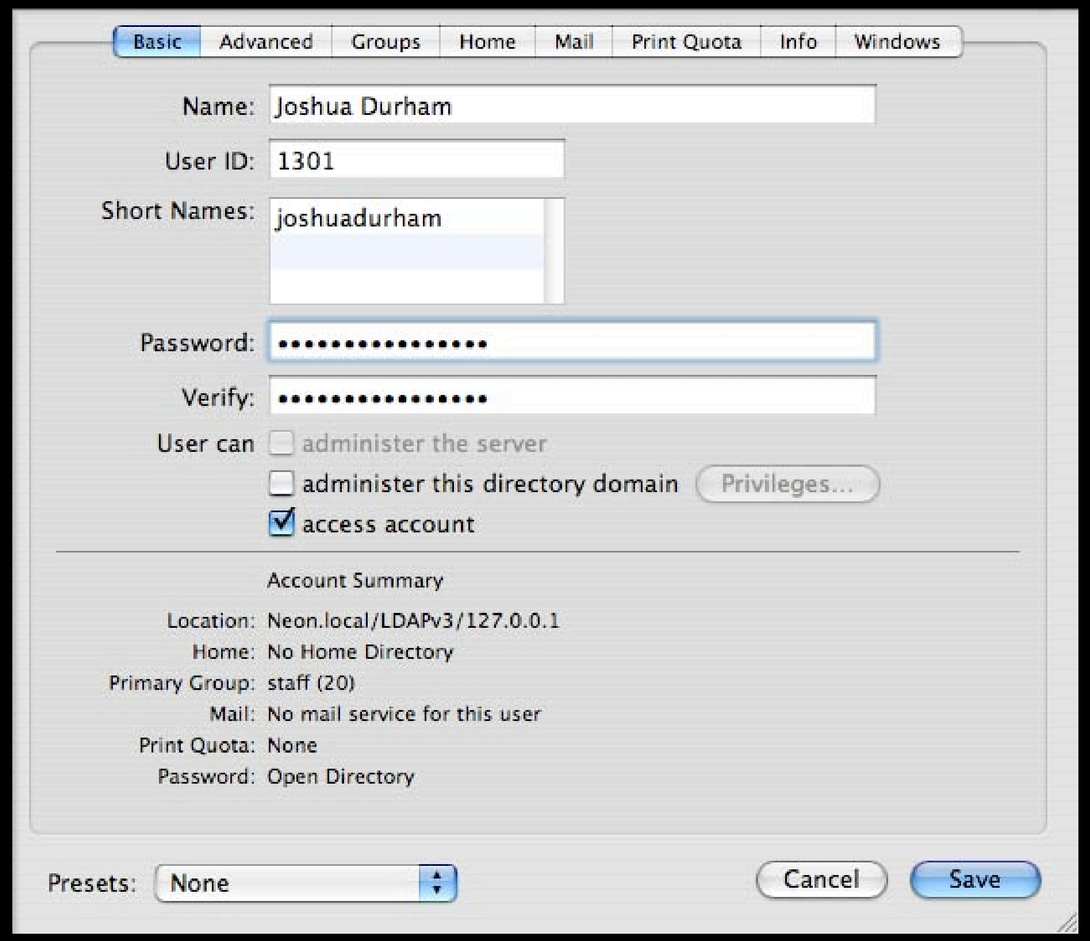
Shared storage (NFS)



- NFS is used as a common store for all user files
- Allows users to run programs across the cluster without having to worry about making sure the necessary files are on each system
- On larger systems, should definitely be given a dedicated system
- Set up using workgroup manager

Service Node Setup

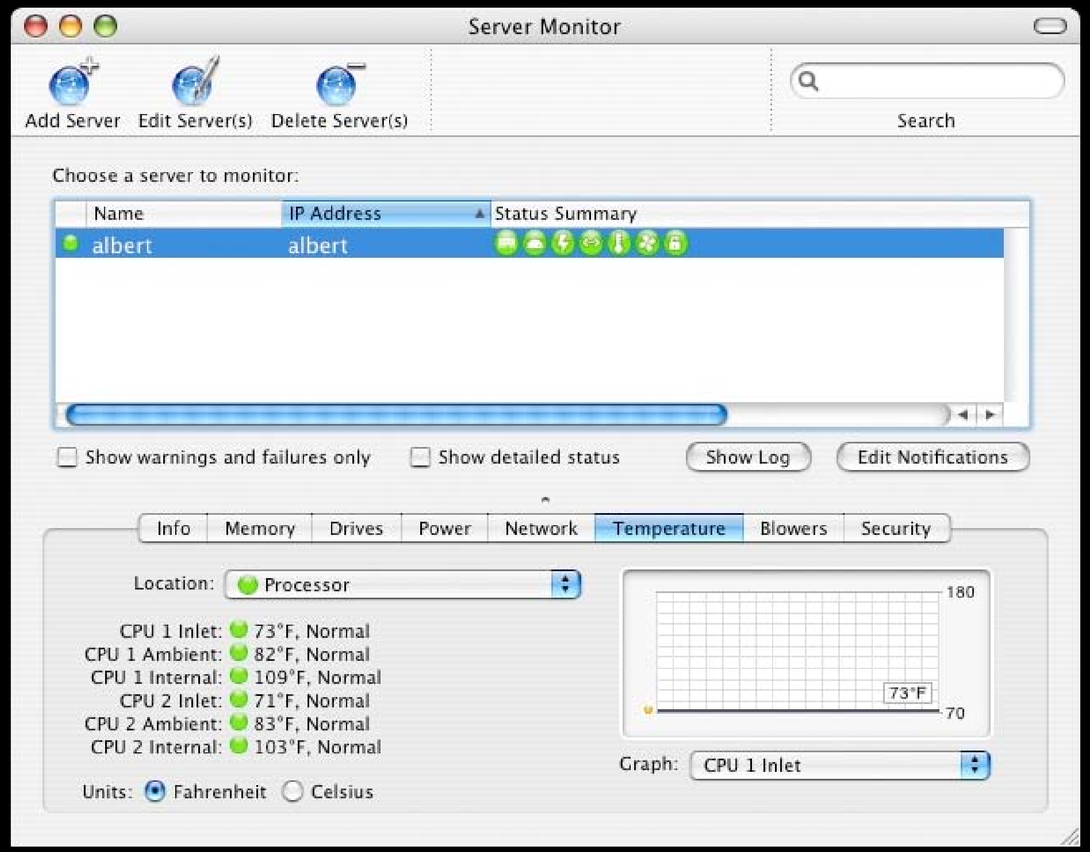
Open Directory (OD)



- OD holds the directory of user accounts on the entire cluster
- Most cluster related jobs require that the user account exists on all the systems in the cluster, with a consistent UID
- Set up using server admin
- Users are maintained using workgroup manager

Service Node Setup

System monitoring



- Allows you to view the hardware status of the cluster nodes from one location
- Can be setup to email on certain events
- Server monitor for monitoring Xserve hardware, Xserve RAIDs can be monitored using Xserve RAID admin

Head-Node Setup

- The head node is the external interface to the cluster
  - The only system that requires to be on the external public network and the internal cluster network
  - Provides remote SSH/SCP access
  - Users often will compile software on this system, and run their codes on the cluster (either directly, or by submitting to a queueing system)
- On small clusters, the head node and service node may be the same

Compute Node Setup

Overview

- A compute node is the ‘workhorse’ of the cluster
  - Should not do anything besides run user codes
- Start with basic image
- Configure with a DHCP IP on en0
- Install ssh key for root user (from service nodes) into authorized_keys
- Point DNS, LDAP, etc., to appropriate service nodes
- Create a compute-node startup script

Compute Node Setup

Basic image for all nodes in the cluster

- Install complete server OS and developer tools
  - Do not enable any services at this point (except maybe ARD client)
- Make sure image has all available, relevant updates applied
- Install any additional software (preferably in/opt)
  - Middleware (MPI, PVM)
  - PDSH (or equivalent)
  - Interconnect software (Myrinet GM, Infiniband VAPI)
  - Compilers and libraries

Compute Node Setup

Startup script example

```
#!/bin/sh

VERSION=1.0
. /etc/rc.common

#
# Mounts the shared NFS volumes at boot time.
#
ConsoleMessage "Mounting Shared Volume from host NFS0"

if [ ! -d /nfs/nfs0 ]; then mkdir -p /nfs/nfs0; fi

mount_nfs nfs0:/Volumes/MirroredRAID /nfs/nfs0
#
# Runs a global "node.startup" script, if one exists.
#
if [ -x /nfs/nfs0/etc/node.start ]; then /nfs/nfs0/etc/node.start; fi
```

Cluster Node Setup

Imaging

- Several methods available for getting software onto the compute nodes
- Most involve using Apple's ASR utility, which is used for copying disk images (DMGs) to a hard drive
- Mike Bombich's NetRestore (http://www.bombich.com) can be used to deploy a DMG to a compute node
- The SneakerNet method: simply take the hard drive of the compute node, put it into an Xserve server, and run ASR locally

Compute Node Setup

After imaging…

- All nodes have a basic core services image
  - All third-party software and tools
- We have a complete service architecture
  - One all-inclusive head node (portal, storage, authentication)
  - Portal node and discrete service nodes
- Cluster nodes
  - Defined in DNS with well-formed host-names
  - Mount a common file store at boot time
  - Are set to run an arbitrary script from the shared store if it exists
  - Trust root from the service nodes to allow password-less authentication

Compute Node Setup

Which allows us to…

- Fix or adjust on “next boot” style management
  - Edit node.start to perform the fix or adjustment on the next reboot
- Spray commands quickly to all nodes in the cluster
  - `pdsh -u 30 -w “node[1-999]” reboot`
- Run basic MPI jobs, including LINPACK

Testing Cluster Components

The Torture Tests

http://developer.apple.com/wwdc2005/

Using LINPACK to Stress-Test Nodes

Run individually on Xserves

- A single two-CPU job per Xserve (P = 1, Q = 2)
- Use a matrix size that maps to about 80% of physical RAM
- Run multiple matrices (four sets is a good start)
  - Compare runs between different Xserve to confirm similar performance
  - First “fix” is always reboot and retry
  - Slower systems may have issues with their logic boards, memory, or CPUs
- Check system logs for ECC errors
  - Replace DIMMs that have high ECC error counts across reboots
- Keep a record of suspect nodes
  - If you lose confidence in a node, replace it

Using LINPACK to Stress-Test Nodes

Run on Xserve pairs

- A single four-CPU job per Xserve pair (P = 2, Q = 2)
- Use a matrix size that maps to about 80% of physical RAM
- Run multiple matrices (four sets is a good start)
  - Compare runs between different pairs to confirm similar performance
  - Slow performance could be a logic board, CPU, memory, or network issue
- Shift pairings (i.e., test nodes 1–2 and 3–4, then 1–3, 2–4, etc.)
  - Helps isolate which node in a given pair is the culprit

Job Scheduling and Resource Management

http://developer.apple.com/wwdc2005/

Resource Management and Job Scheduling

Resource management

- Resource management software runs on compute nodes and gathers data such as system loads
- Often responsible for queuing jobs, executing them, and making sure jobs don’t exceed allocated resources
- Grid engine
  - Free and open-sourced
  - Distributed by Sun
  - Active development community

Resource Management and Job Scheduling

Job scheduling

- Job schedulers work with the resource manager to maximize resource usage
- Helps alleviate the ‘political’ issues of multi-user systems
- Most resource managers come with a basic job scheduler; more complex scheduling is needed if the user base has different agendas
- MAUI
  - Highly optimized and configurable tool capable of supporting a large array of scheduling policies, dynamic priorities, extensive reservations, and fair share
  - Free!

Tweaking the OS For Performance

Less Is More

http://developer.apple.com/wwdc2005/

Tweaking the OS for Performance

Reducing the number of running services

- By reducing the number of running services, you can:
  - Free up CPU cycles for user codes
  - Even if a process does not use a lot of CPU time, it does generate context switches
  - Reduce system complexity, while improving system stability
    - The fewer things running on a system, the fewer things to go wrong
- On the other hand, turning off certain ‘key’ services may make it more difficult to diagnose certain problems
- Should only be done on compute nodes

Tweaking the OS for Performance

Reducing the number of running services

- Mike Bombich created an hpctune script for UIUC's turing cluster
- Runs on a compute node, and goes through and disables certain services
  - `hpctune --recommended` disables a recommended set of services
- Other services can also be specified, such as:
  - `hpctune --recommended --hostconfig AUTOMOUNT NO`
- Being updated for 10.4
- Will be available at http://homepage.mac.com/joshdurham

More Information

http://developer.apple.com/wwdc2005/

Documentation, Sample Code, and Other Resources

http://developer.apple.com/wwdc2005/

Related Sessions

Mac OS X and Scientific Computing	Session 606 Tuesday 3:30PM
HPC Technology Update	Session 609 Tuesday 5:00PM
Monitoring Your System with ARD and Open Source Tools	Session 629 Thursday 2:00PM
Large-scale Xserve RAID Deployment	Session 638 Friday 2:00PM
Enterprise IT Lab	

Who to Contact

Chris Bledsoe
Enterprise IT Segment Manager
cbledsoe@apple.com

Q&A

http://developer.apple.com/wwdc2005/

WWDC2005

Worldwide Developers Conference